Construction of a Moodle-based Placement Test and Possibility of a Moodle-based Computer Adaptive Test

Tetsuo KIMURA
Niigata Seiryo University

Abstract

The present study adds another perspective from item response theory in data analysis to Hinkelman & Grose (2004) which showed the feasibility of using moodle for in-school placement testing, and shows the feasibility and the advantages of using moodle for in-school placement test. In addition, the possibility of the development of a moodle-based computer adaptive test (CAT) is discussed. First, a pilot test consisting of three subtests (vocabulary & grammar, listening comprehension with dialogue, and listening comprehension with monologue) was administered to 268 Japanese students in order to calibrate item difficulties. A Rasch model was utilized to eliminate inappropriate items from the test using fit indices. Each subtest was divided into four 10-minute testlets and administrated separately. After trimming the misfitting items, a 68-item tentative placement test was constructed, which was found to be reliable (KR-20 = .90). Then, three item banks for each subtest were built for the future moodle-based CAT. Finally, item equating steps for the future, and the logic of the future moodle-based CAT were illustrated based on the previous studies in area.

1. Introduction

Because of the high diversity in the English ability of incoming students, many Japanese universities need to administer placement tests to stream their students into levels. Some universities administer commercial placement tests, and some administer in-house placement tests to place their students in appropriate classes. The test is usually given during the freshmen orientation period and its data must be analyzed within a few days. As learning management systems began to spread to Japanese universities, many attempts have been made to apply computer-based placement testing, because the results can be utilized immediately after the administration of the test. In addition, teachers can analyze the items used in the test later and revise them for the future use.

Hinkelman & Grose (2004) conducted a moodle-based pilot placement test for Japanese students and showed the feasibility of using moodle for in-school placement testing. By progressively improving item quality year by year, they concluded that "a self-created placement

test using open source software could, over several years of development, prove equal or superior to generic commercial products in reliability for closed population placement testing"(Hinkelman & Grose, 2004, p. 974). The present study basically followed their procedure and added another perspective from item response theory in analyzing item difficulty and appropriateness. Not only the feasibility and the advantages of using moodle in testing will be shown, but also the possibility of the development of moodle-based computer adaptive tests (CATs) will be discussed.

2. Construction of moodle-based placement test

2.1 Types of questions

The following three types of questions were used for the present study: (1) vocabulary and grammar (Vg), (2) listening comprehension with dialogue (Dlg), and (3) listening comprehension with monologue (Mlg). All the items were adopted from the *Eiken Test Grade pre 1* to *Grade 3*, under the permission of the Society for Testing English Proficiency (STEP): 162 items were used in total (see Table 1 for the details). After all the information necessary to create quizzes on moodle was obtained and organized in an Excel worksheet, it was converted to GIFT files using *Multiple Choice Maker,* so that all the items could be imported to moodle at once. *Multiple Choice Maker*, which was developed by e-learning Service Co., Ltd., is a simple but very helpful Excel file containing a macro to make GIFT files (https://e-learning.ac/moodle-resources/).

Table 1. *Number of Items Used for Each Subtest*

Subtest	Total	Grade pre 1	Grade 2	Grade pre 2	Grade 3
Vg	80	25	20	20	15
Dlg	47	12	15	10	10
Mlg	35	---	15	10	10
Total	162	37	50	40	35

2.2 Pilot test for item analysis

The items in the three subtests (Vg, Dlg, Mlg) were randomly divided into four equivalent testlets on moodle. The entire pilot test consisted of 12 testlets (see Table 2), which were all administered with a time limitation of ten minutes.

In order to gather data to calibrate item difficulty and detect item appropriateness, the pilot test was administered to 268 freshmen in two Japanese universities. The students were told that the test was not compulsory and the results had nothing to do with their grading. Therefore, not all the students answered all the testlets. Sometimes some students stopped answering in the middle of a testlet. Or some students answered all the questions aberrantly, much sooner before they

finished listening to the dialogues or monologues to the end. These aberrant responses to the test were excluded from the data to be analyzed. Thus, the number of persons in each item analysis below is different: 222 in Vg, 142 in Dlg, 119 in Mlg item analysis. Only the data of persons who answered all the items normally in each subtest was used for the item analyses for selecting appropriate items.

Table 2. *Number of Items in each Testlet and its Type of Question*

Testlet	Subtest	Total	Grade pre 1	Grade 2	Grade pre 2	Grade 3
A-1	Vg	20	7	5	5	3
A-2	Dlg	11	3	3	2	3
A-3	Mlg	9	--	4	3	2
B-1	Vg	20	6	5	5	4
B-2	Dlg	12	3	4	3	2
B-3	Mlg	9	--	4	2	3
C-1	Vg	20	6	5	5	4
C-2	Dlg	12	3	4	2	3
C-3	Mlg	9	--	4	3	2
D-1	Vg	20	6	5	5	4
D-2	Dlg	12	3	4	3	2
D-3	Mlg	8	--	3	2	3

2.3 Item analysis and selecting appropriate items

Item analyses were conducted utilizing a Rasch model for each subtest so that inappropriate items could be eliminated from the test and item difficulties could be calibrated according to a standardized logit scale. For the calculation of item difficulty and appropriateness, *TDAP block* was used. *TDAP* (Test Data Analysis Program) was originally developed as an MS-DOS, N88-BASIC program (Ohtomo & Nakamura, 1996), then as a Windows program (Ohtomo, Nakamura & Akiyama, 2002), and now can be used as a block in moodle.

Both items and persons whose fit values expressed in terms of the *t*-distribution exceeding +2.0 were eliminated from the data because values larger than +2.0 "are said to indicate significant departure from the expectations of the model" and "indicate significant misfit" (McNamara 1996, p. 173). The same analysis was repeated until no items nor persons were found to be misfitting.

As for the subtest Vg, item analysis was repeated five times to eliminate all misfit items and persons. The change of the number of items and persons, as well as reliability and mean of the subtest Vg is shown in Figure 1. The reliability of the subtest only improved marginally: the initial

value of KR-20 was .86, and the final .87. The mean was increased from 49.9% to 63.5%, as misfitting items and persons were eliminated. In the end, 44 items out of 80 were eliminated and 36 items remained as appropriate.

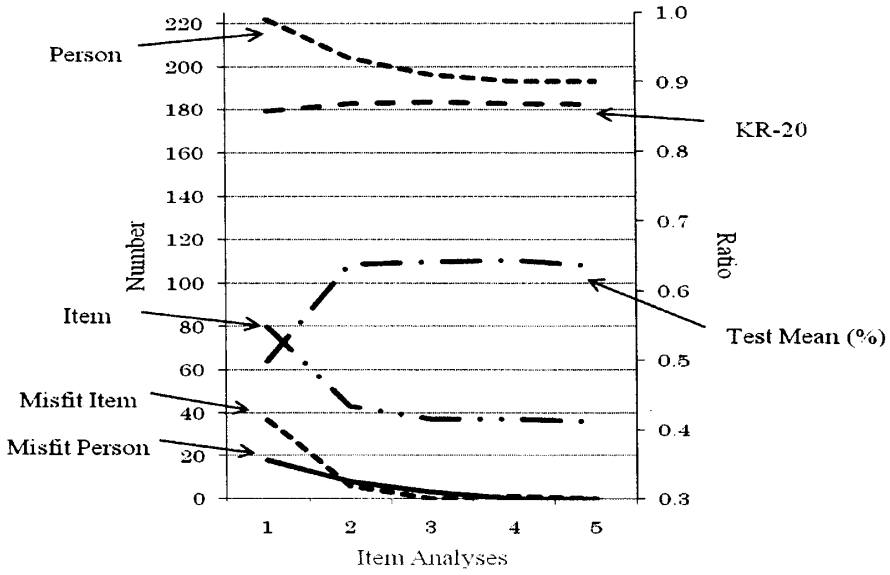

Figure 1. Elimination of misfit items and change of reliability and mean (Vg)

As for the subtest Dlg and Mlg, the same item analyses were repeated nine and three times respectively, and 34 and 16 items were eliminated and 13 and 19 items were remained as appropriate items respectively. The reliability index (KR-20) changed from .72 to .71 for Dlg, and from .75 to .78 for Mlg. The mean values increased from 52.5% to 62.8% and 56.4% to 59.4% respectively. The reliability of the whole test calculated from the subtests' reliability and the correlation coefficient between subtests was .89 at the initial calibration and .90 in the end (see Table 3).

Table 3. *Correlation Coefficient between Subtests and Reliability*

At the First Calibration

	Vg	Dlg	Mlg
Vg	1.00		
Dlg	0.47	1.00	
Mlg	0.59	0.61	1.00
Reliability	0.86	0.72	0.75
Total Reliability		0.89	

At the Last Calibration

	Vg	Dlg	Mlg
Vg	1.00		
Dlg	0.45	1.00	
Mlg	0.64	0.63	1.00
Reliability	0.87	0.71	0.78
Total Reliability		0.90	

2.4 A Tentative placement test

As a result, a tentative placement test was constructed with 36 Vg, 13 Dlg, and 19 Mlg questions. The number of items and basic statistics of the percent correct for each subtest are shown in Table 4. The size of the test seems to be adequate for test administration firstly because it can be done within 45 minutes, and secondly because types of items are well-balanced; about a half of them are listening comprehension and the other half are vocabulary and grammar.

Table 4. *Basic Statistics of Percent Correct in Each Subtest*

Subtest		Total	Grade pre 1	Grade 2	Grade pre 2	Grade 3
Vg	*n*	36	2	10	14	10
	M	64%	30%	51%	65%	80%
	SD	21%	16%	15%	18%	14%
	Max	94%	42%	66%	88%	94%
	Min	18%	19%	18%	29%	52%
Dlg	*n*	13	0	7	2	4
	M	63%	---	45%	82%	84%
	SD	23%	---	13%	9%	11%
	Max	95%	---	66%	91%	95%
	Min	34%	---	34%	73%	70%
Mlg	*n*	19	---	7	5	7
	M	60%	---	51%	52%	74%
	SD	18%	---	12%	18%	15%
	Max	88%	---	64%	73%	88%
	Min	24%	---	28%	24%	50%

When this tentative placement test is administered, the difficulty values (θ) of the items answered correct are summed up as a simplified easy-to-use way to estimate basic ability of English – it can be done only by assigning each difficulty value to each item grade point in a moodle quiz setting. If listening comprehension ability needs to be known separately from basic vocabulary and grammatical ability, it can be estimated as the sum of the difficulty values of the Dlg and Mlg items answered correct. And the basic vocabulary and grammatical ability can be estimated as the sum of the difficulty values of the Vg items answered correct.

However, scores calculated in this way have no meaning on the standardized logit scale calculated in the pilot study. In order to ascertain whether scores are meaningful according to the scores on the scale, the item responses must be analyzed again with the difficulty parameter fixed

as estimated in the pilot test. Because *TDAP block* can not handle that calculation, data must be exported from moodle to another statistical program such as *Easy EstTheta*, a freeware to estimate ability under IRT models (Kumagai, 2005). If the test results are only used to stream students to classes, it is practical to use the scores obtained by this simplified and easy-to-use method. And the placement outcomes will still be the same.

3. Possibility of moodle-based CAT

3.1 Construction of item banks

After the final item difficulties of the pilot test were determined by the last PROX calculation in each subtest analysis, three item banks (IB-VG, IB-DLG, IB-MLG) were built for the future moodle-based CAT. This was done, using *TDAP block*, which has a function to build an item bank based on the item difficulties estimated by PROX. In addition to the item difficulty, a point biserial correlation coefficient as item discrimination power index and actual equivalent number of options are also displayed in the data table of the item bank. And also, the test characteristic curve is displayed along with the data table. The basic statistics of item difficulty (θ) and its *SE* of each item bank are shown in Table 5. Obviously, the number of item in each item bank is not enough for its use on CAT.

Table 5. *Basic Statistics of Item Banks*

Item Bank		*M*	*SD*	*Max*	*Min*
IB-VG (n = 36)	Item difficulty (θ)	-0.707	1.097	1.609	-2.794
	SE of θ	0.160	0.034	0.267	0.135
IB-DLG (n = 13)	Item difficulty (θ)	-0.644	1.202	0.743	-2.782
	SE of θ	0.180	0.050	0.313	0.147
IB-MLG (n = 19)	Item difficulty (θ)	-0.455	0.875	1.221	-1.931
	SE of θ	0.188	0.024	0.244	0.171

3.2 Further pilot tests for data collection and equating

The tentative placement test, in which the items are fixed as shown in Table 6, will be administrated next year to incoming students, and the additional pilot tests (12 testlets: E-1 to H-3, see Table 6) will be assigned as homework or in-class assignments. Both placement and pilot tests will be conducted on moodle because students, as well as teachers, can get the results of the test immediately, and also because the item responses can be easily retrieved. The results of the additional pilot tests will be analyzed by *TDAP block* in moodle. After trimming inadequate items from the data using misfit statistics as before, the final values of item difficulty will be equated with the anchor items (i.e. items in the placement test) and added to each item bank.

After a few years of practice of this procedure, a considerably large number of adequate items will be saved in to each item bank. When about 200 items are pooled in each item bank, it will be time to develop a moodle-based CAT.

Table 6: *Tentative placement test and new testlets*

Testlet	Type	Total	Grade pre 1	Grade 2	Grade pre 2	Grade 3
Tentative	Vg	36	2	10	14	10
Placement Test	Dlg	13	0	7	2	4
(Anchor Items)	Mlg	19	---	7	5	7
E-1	Vg	20	7	5	5	3
E-2	Dlg	11	3	3	2	3
E-3	Mlg	9	--	4	3	2
F-1	Vg	20	6	5	5	4
F-2	Dlg	12	3	4	3	2
F-3	Mlg	9	--	4	2	3
G-1	Vg	20	6	5	5	4
G-2	Dlg	12	3	4	2	3
G-3	Mlg	9	--	4	3	2
H-1	Vg	20	6	5	5	4
H-2	Dlg	12	3	4	3	2
H-3	Mlg	8	--	3	2	3

3.3 A moodle-based CAT

Wainer and Kiely (1987) suggest the use of testlets in CAT rather than individual items because context effects can be reduced, and because it is easier for test developers to create good testlets for a CAT than to create good items. They claimed that at least some CATs are best constructed and scored as a set of testlets rather than as a set of a larger number of individual items. And further, Thissen et al. (1989) proved that the resulting testlet scores show higher validity than scores derived at the item level. Other studies further proclaimed that although multistage testing had been neglected in favor of the consideration of individual-item-level CATs, "both the computers and the IRT models have advanced to the point at which testlet-CATs are practical" (Thissen & Mislevy, 2000, p. 128).

In future studies, several testlets which have different levels of difficulty will be created for each subtest (Vg, Dlg, Mlg) and a moodle-based CAT will be administrated using *CAT module,* which was originally developed by E-leaning Service Co., Ltd. for a Japanese language placement test for foreign students in Japan. It is now under revision and being updated for the latest version

of moodle. For the details of the item bank system developed within the framework of moodle, see Ito (2008). In this pilot phase of the study, many things, such as the optimal number of items in a testlet, how to decide the first testlet, how to select the next optimal testlet, and where to stop the CAT, need to be investigated. However, the feasibility and practicality, as well as the possibility of developing a moodle-based CAT can be ascertained in the future.

4. Conclusion

Chapelle (2000) lists six qualities to consider in evaluating the usefulness of a CAT: reliability, construct validity, authenticity, interactiveness, positive impact, and practicality. This study only showed the possibility of developing moodle-based CAT, focusing only on reliability and practicality. Although the cases are limited in number, the present study showed that test administration and item analyses on moodle are feasible and practical, and that the moodle-based in-house placement test, which is still not adaptive but item-fixed, is reliable and practical. In addition, it is anticipated that a future moodle-based CAT will provide a practical, reliable platform for administering placement tests to large numbers of students in a short time.

The four other qualities Chapelle listed: construct validity, authenticity, interactiveness, and positive impact of CAT still remain unanswered. Further study to investigate these qualities is necessary, and the present results and implications are necessary to be confirmed with much larger amount of data. And the model to analyze test data, as well as the algorithm, may be reconsidered in future studies.

References

Chapelle, C. (2000). *Computer applications in second language acquisition: Foundations for teaching, testing, and research.* Cambridge, U.K.: Cambridge University Press.

Hinkelman, D., & Grose, T. (2004). Placement testing and audio quiz-making with open source software. *Proceedings of CLaSIC 2004*, 972-981.

Ito, S. (2008). *Ryugakusei no Nihongo Nouryoku Sokutei no tameno Test Koumoku Pool no Kouchiku. [Construction of Item Pool for the Assessment of Foreign Students' Japanese Ability].* Kagaku Kenkyhuhi Hojokin Kennkyuu Seika Houkokusho [Working Paper of Grant-in-aid Scientific Research] 16202008.

Kumagai, R. (2005). Easy EstTheta Ver0.1.1 [Computer software]. Retrieved July 31, 2007, from http://itranalysis.main.jp

McNamara, T. F. (1996). *Measuring Second Language Performance.* Essex: Addison Wesley Longman Limited.

Ohtomo, K., & Nakamura, Y. (1996). Test Data Analysis Program (TDAP) Ver. 1.0 [MS-DOS, N88-BASIC, Computer software]. In K. Ohtomo, *Kohmoku Ohto Riron Nyumon*

[Introduction to Item Response Theory]. Tokyo: Taishukan.

Ohtomo, K., Nakamura, Y., & Akiyama, M. (2002). Test Data Analysis Program (TDAP) Ver. 2.0 [Windows, Computer software]. In K. Ohtomo (ed.) & Y. Nakamaura, *Test de Gengo Nouryoku ha Hakarerunoka: Gengo Test Data Bunseki Nyumon [Can Test Assess Language Ability? Introduction to Data Analysis of Language Test].* Tokyo: Kiriharashoten.

Thissen, D., Steinberf, I., & Mooney, J.A. (1989). Trace lines for testlets: A use of multiple categorical-response models. *Journal of Educational Measurement, 26,* 247-260.

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In Wainer, H. (with Dorans, N.J., Eignor, D., Flaugher, R., Green, B.F., Mislevy, R.J., Steinberg, L., and Thissen, D.), *Computerized Adaptive Testing A Primer* (2nd ed., pp. 101-133). NJ: Lawrence Erlbaum Associates.

Wainer, H., & Kiely, G.L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185-201.